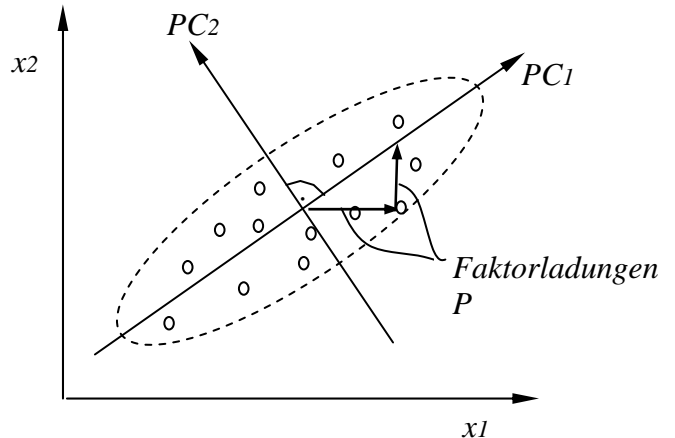


Hauptkomponentenanalyse

Principal Component Analysis PCA

Die Hauptkomponentenanalyse berechnet aus ursprünglichen Daten (Variablen x) neue so genannte latente Variablen, die man kurz als Faktoren bezeichnet und die die **Hauptkomponenten** PC darstellen. Ziel ist es mit wenigen Faktoren alle Ausgangsvariablen zu beschreiben (Datenreduktion). Anhand von 2 Ausgangsvariablen x_1 und x_2 und deren Messpunkte, soll das Prinzip wie rechts im Bild dargestellt werden. Die Messpunkte liegen in einer Ellipse, die in Ihrer Form und Neigung von der Korrelation zwischen den Variablen abhängen.



Durch Verschiebung des Nullpunktes und Drehung des Koordinatensystems entsteht ein neues Achsensystem. Die erste so genannte Hauptachse weist in Richtung der größten Streuung der standardisierten Werte von x_1 und x_2 . Die zweite Hauptachse steht senkrecht auf der ersten, wobei hier der nächst geringere Varianzanteil erklärt wird. Die Hauptkomponenten bezeichnet man deshalb auch als Eigenvektoren.

Zur Bestimmung von Hauptkomponenten werden so genannte Faktorladungen P (Loadings) und Score-Werte T gebildet. Die Faktorladungen ergeben auf dem ursprünglichen Koordinatensystem von x_1 und x_2 die Lage der PC . Die Dimension der Faktorladungen ist Anzahl Komponenten \times Anzahl Variable x . Die Score-Werte T beschreiben die Projektionen auf die Hauptachsen für jeden Punkt. Die Dimension von T ist Anzahl Komponente \times Anzahl Messwerte. In Matrixschreibweise ist der Zusammenhang:

$$X = T P^T$$

Für die Faktorladungen gilt die Bedingung:

$$p_1^2 + p_2^2 + \dots + p_k^2 = 1$$

Die Hauptkomponenten werden über die Score-Werte t_i und die so genannten Eigenwerte λ_i berechnet:

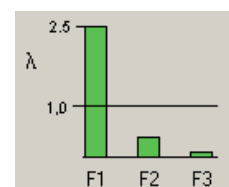
$$PC_i = \frac{t_i}{\sqrt{\lambda_i}}$$

Der Eigenwert λ_i gibt an, wie viel von der Gesamtvarianz aller Variablen durch diesen Faktor erfasst wird. Die Eigenwerte dienen auch zur Entscheidung, ob Faktoren im Modell beibehalten oder weggelassen werden können. Ist der Eigenwert kleiner oder gleich 1, erklärt er weniger oder nur gleich viel wie die Varianz einer einzigen Variablen. Damit kann der Faktor weggelassen werden. Eigenwerte und Eigenvektoren ergeben zusammen eine voneinander unabhängige (orthogonale) Struktur.

Die Eigenwerte lassen sich nicht analytisch berechnen, sondern müssen iterativ bestimmt werden (Eigenwertproblem). Für weitere Details sei auf die einschlägige Literatur verwiesen (insbesondere *Multivariate Datenanalyse* /18/).

Beispiel: Gegeben sind die Variablen x_1 , x_2 und x_3 . Berechnet wurde der Faktor F :

x_1	x_2	x_3	F
1	3	4	-1,00
2	4	3	-0,70
3	1	1	1,00
4	2	2	0,70



Für diese Daten reicht ein Faktor aus (λ für den zweiten und dritten Faktor ist unter 1). Neben den Faktorladungen gibt es noch die so genannten Korrelations-Ladungen. Das sind die Korrelationen zwischen den Faktoren und den ursprünglichen Variablen. Betrachtet man die Korrelationen zueinander, so zeigt sich, dass der neue Faktor mit allen Ausgangsvariablen deutlich höher korreliert, als die Variablen untereinander. Ziel des Faktors ist es ja gerade eine möglichst gute „Beschreibung“ aller Variablen gemeinsam zu erreichen.

$x_3: F$	-0,958
$x_2: F$	-0,881
$x_1: F$	0,881
$x_1: x_3$	-0,800
$x_2: x_3$	0,800
$x_1: x_2$	0,600

Korrelationen zwischen Faktor und Ausgangsvariablen

Korrelations-Ladungen

Zu beachten ist hier für die Interpretation, dass der Faktor mit Variable x_2 und x_3 negativ und mit Variable x_1 positiv korreliert. Es bedeutet eine negative Korrelation, dass für die Veränderung der Faktorwerte in einer Untersuchung von kleine auf große Werte die Richtung für Variable x_3 und x_2 umgekehrt ist.

Partial Least Square (PLS)

PLS wurde 1960 von dem schwedischen Ökonometriker Herman Wold entwickelt /13/, /14/. PLS steht für: „*Partial Least Squares Modeling in Latent Variables*“. Der Zweck ist vor allem die Auswertung von korrelierenden Daten oder von Mischungsplänen, bei denen die Multiple Lineare Regression (MLR) nicht anwendbar ist. Ein wesentlicher Vorteil von PLS ist auch, dass viele Variable verarbeitet werden können. Es ist sogar möglich mit weniger Versuchen auszukommen, als Variable vorhanden sind. Dies ist mit MLR in der Form nicht möglich.

PLS ist mit der Hauptkomponentenanalyse PCA (*Principle Component Analysis*) sehr verwandt. Im Gegensatz zu PCA gilt hier der Zusammenhang mit der Gewichtsmatrix W anstelle der Loadings:

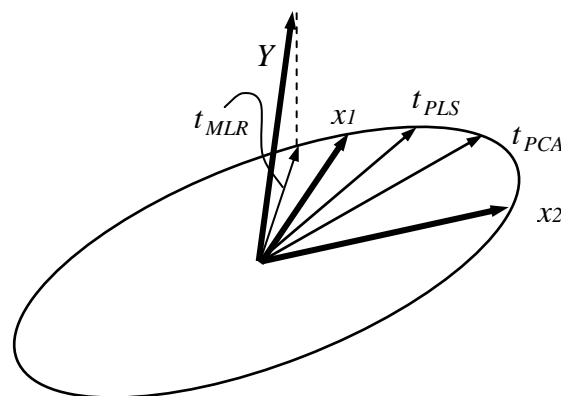
$$X = TW^T$$

In W ist die Zielgröße y enthalten, die es in der PCA nicht gibt. Auch hier gilt $w_1^2 + w_2^2 + \dots w_k^2 = 1$

Das Regressionsmodell ist:

$$\hat{y} = Tc^T$$

Wobei c die Regressionskoeffizienten darstellt. Das dargestellte Bild zeigt zwei Variablen x_1 und x_2 . Die Hauptkomponentenanalyse PCA mit t_{PCA} liegt in der „Beule“ der Ellipse, die sich in der Ebene von x ergibt. Je stärker x_1 und x_2 korrelieren, desto länger wird t_{PCA} . Liegt keine



Korrelation vor, ist die Vektorrichtung von t_{PCA} nicht mehr definiert, denn die Ellipse wird dann zu einem Kreis und hat keine Vorzugsrichtung mehr. Die Komponente t_{PLS} dagegen ist über die Betrachtung der Kovarianz dann immer noch bestimmbar. Das ist ein

entscheidender Vorteil von PLS gegenüber PCA. Die Ergebnisse, d.h. die ermittelten Koeffizienten der Variablen sind dann identisch mit der MLR-Methode (für orthogonale Daten). Während die MLR-Methode bei hochkorrelierenden Daten nicht mehr eindeutige Ergebnisse liefert oder ganz aussteigt, kann die PLS-Methode weiterhin angewendet werden. Selbst wenn zwei Variablen zu 100% korrelierend ist das noch möglich. Natürlich ist die Zuordnung der Effekte dann nicht mehr eindeutig, PLS vergibt in diesem Fall den beiden Variablen jeweils den halben Effekt zu gleichen Anteilen.

Der Nachteil des PLS-Verfahrens ist, dass die Prognosen und R^2 schlechter sind als bei MLR. Auch sind die Koeffizienten teilweise wesentlich kleiner, was dazu führt die Effekte zu gering zu schätzen.

Der komplette Algorithmus (NIPALS – *Nonlinear Iterative Partial Least Square*) stellt sich wie folgt dar:

$w' = \frac{X^T y}{y^T y}$	<i>Wichtungen absolut für standardisierte Matrix X</i>
$w = w' / \sum w'^2$	<i>Wichtungen normiert</i>
$t = Xw$	<i>Score Vektor</i>
$= \frac{\sum_{j=1}^z \text{cov}(y, x_j) x_j}{\sum_{j=1}^z \text{cov}(y, x_j)^2}$	<i>mit z = Anzahl Variable</i>
$c = \frac{y^T t}{t^T t}$	<i>Regressionskoeffizienten zwischen y und Komponente</i>
$p = \frac{X^T t}{t^T t}$	<i>Ladungs-Vektor</i>
$E = X - tp^T$	<i>Residuen-Matrix der Variablen</i>
$f = y - tc^T$	<i>Residuen-Vektor der Zielgröße</i>

Die nächsten Komponenten werden bestimmt, indem man $X=E$ und $y=f$ setzt und von vorne berechnet. Bezogen auf den Regressionsansatz zwischen y und den ursprünglichen Variablen errechnen sich die Koeffizienten b über:

$$b = W(P^T W)^{-1} c^T$$

Zusammenfassende Eigenschaften:

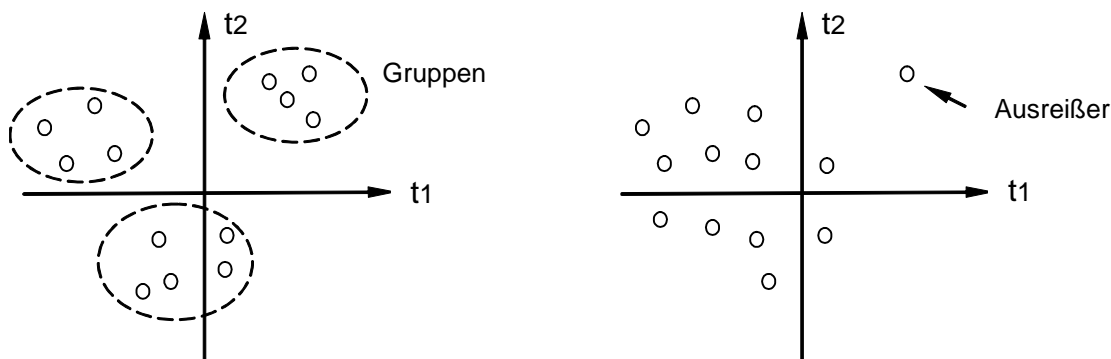
- R^2_{PLS} ist kleiner R^2_{MLR}
- Koeffizienten PLS sind kleiner als bei MLR -> Fehler wirken sich dadurch auch geringer aus.
- PLS maximiert die Kovarianz zwischen Hauptkomponenten und Y , MLR maximiert dagegen die Korrelation
- PLS kann mit hohen Korrelationen zwischen den Variablen X umgehen.

PLS hat sich in den Bereichen Pharma, Chemie und Spektroskopie als Standard durchgesetzt. Häufig wird es Universalmethode für alle Auswertungen gesehen. Für Auswertung von Daten die nicht zu stark oder gar nicht korrelieren (z.B. aus der Versuchspla-

nung) ist aber nach wie vor die Multiple Regression vorzuziehen, da hier die Effekte und Modelle besser zu interpretieren sind. Bei rein orthogonalen Daten sind allerdings die Koeffizienten der Regressionsmodelle auch gleich.

Score Plot

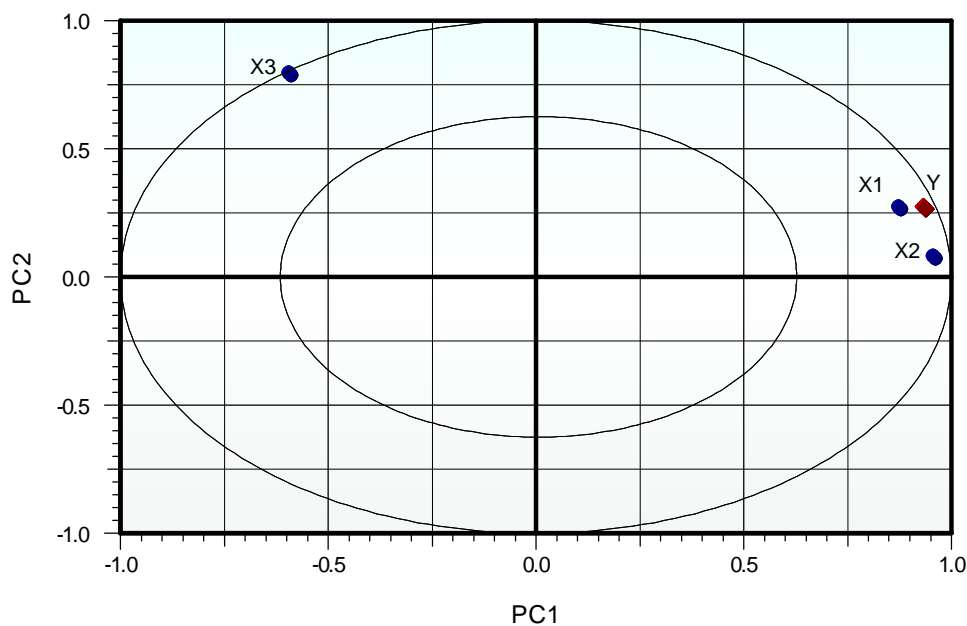
Der Score Plot stellt jeden Messpunkt über die wichtigsten Scores t_1 und t_2 dar. Dabei können evtl. Muster und Gemeinsamkeiten erkannt werden. Liegen Messpunkte eng beieinander, desto ähnlicher sind sie sich und umgekehrt. Dabei können auch markante Ausreißer erkannt werden.



Korrelations Ladungen (Correlation Loading Plot)

Im so genannten Correlation Loading Plot werden die erklärten Varianzen der Variablen und der Zielgröße auf die Komponenten PC dargestellt. Die Achsen sind skaliert als Korrelationswerte wobei gilt: Erkl. Varianz = Korrelation^2 . In diesem Diagramm werden die Einflüsse der Variablen aufgezeigt und man erkennt welche Komponenten besser die Variablen beschreiben.

Die Ellipsen stellen jeweils 100% (äußere) und 50% (innere) erklärte Varianz dar.



Je näher die Variablen an der 100% Ellipse liegen, desto wichtiger sind diese. In diesem Beispiel erklärt die Komponente PC1 die Variablen x_1 , x_2 , sowie die Zeilegröße fast alleine, während für x_3 die beide Komponenten notwendig sind.

Schätzung der Streuung

Die Streuung der Koeffizienten b kann hier nicht grundsätzlich, wie bei der MLR-Methode aus der Spur von $(X^T X)^{-1}$ berechnet werden. Ist die Korrelation zwischen den Variablen groß, so kann eine Streuung nur über eine so genannte Kreuzvalidierung erfolgen. Die häufigst verwendete Variante lässt sich durch folgendes Verfahren beschreiben. Jede Zeile aus der Matrix X wird einmalig herausgelassen (one leave out). Mit der reduzierten Matrix wird das Modell mit seinen Koeffizienten mit PLS berechnet. Es ergeben sich somit n unterschiedliche b . Hieraus kann man eine Standardabweichung für b bestimmen und somit die Streuung schätzen. Dabei wird klar, dass die ermittelte Standardabweichung stark von der Anzahl der Versuche abhängig ist, was bei geringem Umfang kritisch sein kann. Auch liefern Varianten der Kreuzvalidierung, z.B. bei Herauslassen von gleichzeitig mehreren Zeilen mit unterschiedlichen Umfängen, unterschiedliche Werte. Die Berechnung und Anwendung der p-Values, wie beim MLR-Verfahren ist hier also nicht zu empfehlen. Deshalb finden man für das PLS-Methode stattdessen häufig die *VIP*-Kennzahl zur Variablenselektion.

Variablenselektion mit VIP

Für das PLS-Verfahren eignet sich zur Variablenselektion die *VIP*-Kennzahl. *VIP* steht für *Variable Importance in the Projection*, also wie wichtig der Einfluss der Variable in der Projektion auf die Scores t ist. Diese Kennzahl wurde erstmals 1993 von Wold /13/ veröffentlicht. *VIP* berechnet sich für die jeweilige Variable x_j über:

$$VIP_j = \sqrt{z \sum_{k=1}^h \left(\frac{y^T t_k}{t_k^T t_k} w_{jk}^2 \right) / \sum_{k=1}^h \left(\frac{y^T t_k}{t_k^T t_k} \right)} \quad \begin{array}{l} \text{mit } h = \text{Anzahl Komponenten,} \\ z = \text{Anzahl Ausgangsvariable (bzw. Terme)} \end{array}$$

Der y -Vektor muss hier standardisiert sein. In der Literatur wird für die *VIP*-Zahl ein Grenzwert zwischen 0,8 ...1 genannt. Zu kleine Werte bedeuten, dass die Variablen aus dem Modell weggelassen werden sollten. Die Erfahrungen zeigen jedoch, dass auch *VIP*-Werte unter 0,5 nicht ungewöhnlich sind für Variable, die von Ihrem Einfluss auf das Modell wichtig sind. Bei der Frage, ob eine Variable aus dem Modell genommen werden sollte, ist also auch die Größe des Einflusses (Koeffizient) zu berücksichtigen. Letztlich gilt auch hier, dass die Frage ob ein Term im Modell bleiben soll, nicht nur über eine Kennzahl (p-Value bei MLR, *VIP* bei PLS) entschieden werden sollte. Es besteht auch die Frage nach der Größe der Koeffizienten und das Wissen der tatsächlichen physikalischen Zusammenhänge.

Neuere Verfahren O-PLS

Ein neueres Verfahren ist das so geannte O-PLS (Orthogonal Partielle-Least-Square). Hier wird nur der Anteil der X-Richtungen auf die Komponente berücksichtigt. Die Interpretation der Komponenten und die Korrelation zu Y wird damit besser. Die zurückgerechneten Koeffizienten b auf die ursprünglichen Variablen und die Modellprognosen ändern sich jedoch nicht. Der praktische Vorteil für O-PLS liegt alleine in der Behandlung und Erklärung der Komponenten. Bleibt man bei der Betrachtungsweise zwischen den Erklärungsanteilen der Variablen und der Zielgröße, so bringt O-PLS keinen Vorteil.